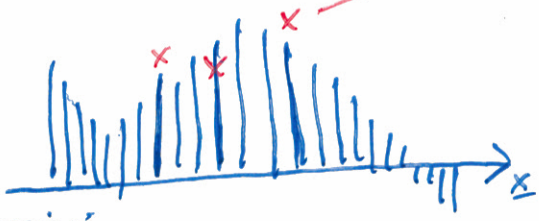
Gaussian Processes



Noisy observations of a subset values.

- Can infer any subset of function values given these.

Gaussian:

Prior directly on the function values at all possible inputs

Marginal $p(\underline{f}) = N(\underline{f}; \underline{0}, K)$

$\underline{f}$: vector of values at locations $\{\underline{x}^{(n)}\}$

K: $K_{ij} = k(\underline{x}^{(i)}, \underline{x}^{(j)})$

$= \text{cov}[f_i, f_j]$

GPs are Bayesian Linear Regression

Model

$$f_i = f(\underline{x}^{(i)}) = \underline{w}^T \underline{x}^{(i)} + b$$

Prior: $\underline{w} \sim N(0, \sigma_w^2 I)$, $b \sim N(0, \sigma_b^2)$

$\left(\sigma_{prior}, \frac{1}{\alpha}\right)$

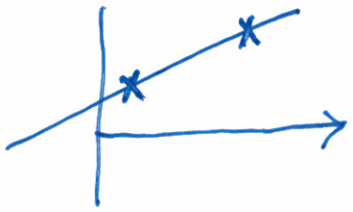
$$\text{cov}(f_i, f_j) = \mathbb{E}[f_i f_j] - \underbrace{\mathbb{E}[f_i]}_{0}\underbrace{\mathbb{E}[f_j]}_{0}$$

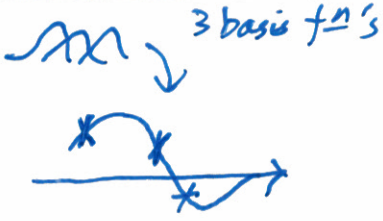
$$= \mathbb{E}\left[\left(\underbrace{\underline{x}^{(i)T}\underline{w}}_{\underline{w}^T\underline{x}^{(i)}} + b\right)\left(\underline{w}^T\underline{x}^{(j)} + b\right)\right]$$

$$= \mathbb{E}\left[\underline{x}^{(i)T}\underline{w}\,\underline{w}^T\underline{x}^{(j)} + b^2 + \dots\right]$$

$$= \underline{x}^{(i)T}\underbrace{\mathbb{E}[\underline{w}\underline{w}^T]}_{\text{cov}(\underline{w})}\underline{x}^{(j)} + \mathbb{E}[b^2] + \underbrace{\dots}_{0}$$

$$= \sigma_w^2\, \underline{x}^{(i)T}\underline{x}^{(j)} + \sigma_b^2 = k(\underline{x}^{(i)}, \underline{x}^{(j)})$$





Basis Functions

$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_w^2 \underline{\phi}(\underline{x}^{(i)})^T \underline{\phi}(\underline{x}^{(j)}) + \sigma_b^2$$

We only need inner products of features.

"Kernel trick"

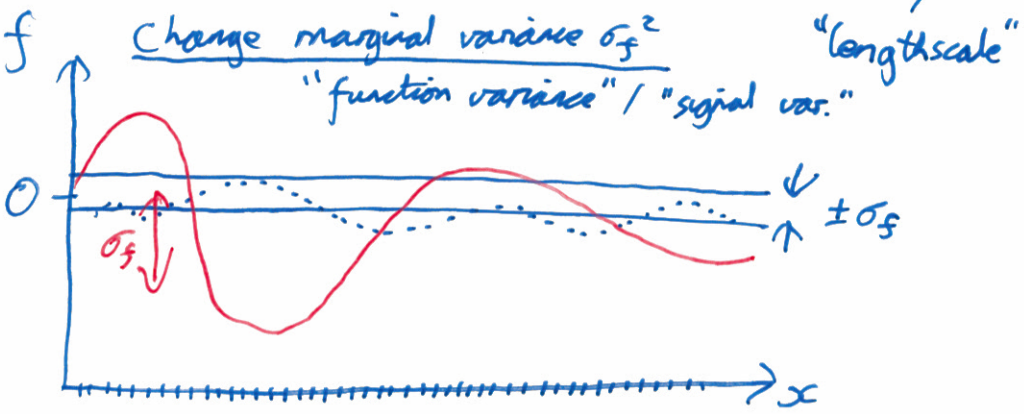
- Rewrite algorithm so it only needs inner products of features.
- We then use very large / infinite set of basis functions.
- Replace that inner product with analytic expression we can compute.

It can be shown that....

If we put RBFs <u>everywhere</u> we can derive a kernel:

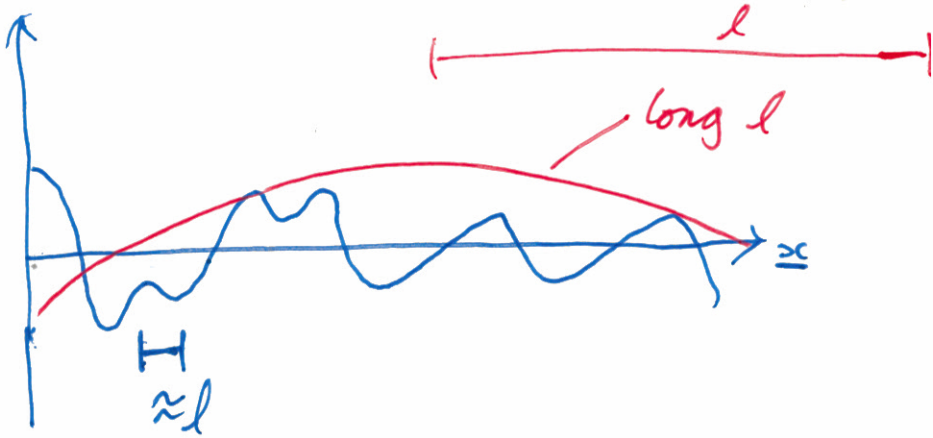
$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_d \left(x_d^{(i)} - x_d^{(j)}\right)^2 / \ell_d^2\right)$$

"lengthscale"



$k(\underline{x}^{(i)}, \underline{x}^{(i)}) = \sigma_f^2$ Diagonal of cov. matrix

Change lengthscale



Pick parameters by (marginal) likelihood.

$$p(y \mid X, \theta = \{\sigma_s^2, \ell_d, \sigma_n^2, \ldots\}) = N(y; 0, K + \sigma_n^2 I).$$

↑

$O(N^2)$ RAM

Inverting / Factoring costs $O(N^3)$...